KLASTERISASI PADA DOKUMEN BERITA BERBAHASA INDONESIA BERDASARKAN FREQUENT TERM-BASED TEXT CLUSTERING (HFTC DAN FTC)

Mega Rulliana[1], M. Arif Bijaksana[2], Angelina Prima Kurniati[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Berkembangnya teknologi di dunia maya membuat jumlah informasi berupa artikel berita semakin banyak. Untuk itu, diperlukan suatu pengelompokan terhadap artikel yang memudahkan pembaca mencari informasi dengan menerapkan salah satu fungsionalitas dari data mining, yaitu klasterisasi. Teknik klasterisasi yang ada saat ini masih belum secara tepat menangani data berdimensi tinggi dan database yang berukuran besar sehingga deskripsi dari klaster tersebut masih sulit untuk pahami. Oleh karena itu dibutuhkan metode pengklasteran dimana hasil pengklasteran tersebut memiliki bentuk deskripsi klaster yang mudah dipahami.

Metode yang dapat diterapkan ini mampu mengurangi dimensionalitas data yang tinggi dan besarnya ukuran database. Ada beberapa metode yang dapat digunakan yaitu berdasarkan frequent term-based text clustering yang terdiri dari hierarchical frequent term-based clustering (HFTC) dan frequent term-based clustering (FTC).

Hasil dari klasterisasi berdasarkan frequent term based text clustering adalah berupa klaster yang memiliki deskripsi klaster yang mudah dipahami. Berdasarkan hasil percobaan dapat disimpulkan bahwa pada HFTC, F-measure nilainya semakin besar dengan minimum support yang semakin kecil. Hal ini menunjukkan kualitas klaster yang terbentuk pun semakin bagus. Nilai Entropy yang dihasilkan pada FTC bervariasi dan tidak memiliki pola pada tiap minimum support yang diinputkan. Serta Waktu yang dibutuhkan dalam pembentukan klaster akan semakin sedikit seiring dengan makin besarnya nilai minimum support.

Kata Kunci : klasterisasi, frequent term-based text clustering, HFTC, FTC, Fmeasure, Entropy

Abstract

The development of large numbers of information like news articles are available on the internet. Hence text clustering is needed by applying clasterisation as one of data mining task. Nowdays, the method of text clustering still do not really address the special problem of text clustering such as the high dimensionality of the data and very large size of the database, therefore understandability of the cluster description still difficult to understand.

This application method can reduce the high dimensionality of the data and very large size of the database. There is some methods that can be used based on frequent term-based text clustering, such as hierarchical frequent term-based clustering (HFTC) and frequent term-based clustering (FTC).

The clusterisation's output that based on frequent term-based text clustering has the understandability of the cluster description. Based on experimental evaluation, it can be concluded on HFTC, f-measure value increasing while minsup decreasing thus the quality of cluster is better, on each minimum support, FTC has variation Entropy value, and the time to make cluster is decreasing while minimum support is increasing.

Keywords : clusterisation, frequent term-based text clustering, HFTC, FTC, f-measure, Entropy

1. Pendahuluan

1.1 Latar belakang

Penggunaan Internet saat ini telah memacu pertumbuhan dan pertukaran informasi yang sangat pesat dibandingkan tahun-tahun sebelumnya. Jumlah informasi pun terus meningkat secara eksponensial. Hal ini seiring pula dengan jumlah sumber berita berbahasa Indonesia di internet yang semakin besar. Karena jumlah yang semakin besar inilah pengguna internet kesulitan untuk mendapatkan dokumen berita yang diinginkan. Untuk itu diperlukan pengelompokkan dokumen berita berbahasa Indonesia berdasarkan informasi yang terkandung di dalamnya. Sehingga dokumen berita tersebut bisa dikelompokkan pada topik tertentu.

Pengelompokan dokumen berita berdasarkan kesamaan isi untuk pencarian dan penggunaan informasi berita sangatlah penting untuk memudahkan pengguna dalam mencari informasi yang diinginkan. Untuk pengelompokkan dokumen dimana class label dokumennya belum ditentukan dapat dilakukan dengan menggunakan teknik klasterisasi. Saat ini telah banyak dikembangkan teknik klasterisasi, namun teknik klasterisasi yang ada saat ini masih belum secara tepat menangani data berdimensi tinggi sehingga deskripsi dari klaster tersebut masih sulit untuk dimengerti. Oleh karena itu dibutuhkan metode pengklasteran dimana hasil pengklasteran tersebut memiliki bentuk deskripsi klaster yang mudah dipahami.

Salah satu cara untuk mendapatkan hasil klasterisasi dengan deskripsi klaster yang mudah dipahami adalah dengan melakukan pengurangan dimensi pada *term* (kata hasil preprosesing). Dan, metode yang dapat mengklasterkan dokumen berita dengan terlebih dahulu melakukan pengurangan dimensi dikenal dengan pendekatan berdasarkan *Frequent Term-based Text Clustering*, dimana *Frequent Term-based Text Clustering* mengklasterkan dokumen berdasarkan *frequent termset* atau *term* yang sering muncul dibanyak dokumen. Salah satu model pengaplikasiannya antara lain *Frequent Term-Based Clustering* (FTC) dimana hasil klaster dari FTC ini bersifat non-overlapping atau tidak saling beririsan tiap dokumennya antar klaster. Dengan kata lain satu dokumen hanya terdapat tepat di satu klaster saja. Metode lainnya selain FTC yaitu *Hierarchical Frequent Term-Based Clustering* (HFTC) dimana hasil klaster dari HFTC ini berupa klaster-klaster yang dokumen di dalamnya saling overlap atau beririsan. Overlap disini maksudnya adalah pada klaster yang berbeda terdapat dokumen yang sama. Selain itu deskripsi klaster pada FTC dan HFTC lebih mudah dipahami.

1.2 Perumusan masalah

Dalam Tugas Akhir ini penulis merumuskan beberapa masalah yang timbul dari latar belakang masalah yang dipaparkan diatas, yaitu antara lain:

- Bagaimana cara memproses kata pada dokumen menjadi *term*.
- Bagaimana cara mengurangi *term* yang tidak *frequent*.
- Bagaimana cara mengelompokkan dokumen berita dengan menggunakan metode klasterisasi, dimana deskripsi klaster tersebut mudah dimengerti.

Hipotesa awal dari perumusan masalah di atas adalah :

1 Pemilihan sejumlah *frequent termset* dapat dengan menggunakan algoritma *apriori*.
2 Untuk melakukan pengelompokan dokumen berdasarkan *term* dapat dilakukan dengan menggunakan algoritma *HFTC* dan FTC dimana keduanya menjamin kecilnya dimensi hasil pengklasteran, sehingga deskripsinya mudah dimengerti.
3 Untuk melakukan analisis kualitas klaster hasil pengklasteran dapat dengan menggunakan penghitungan *F-Measure* untuk HFTC dan *Entropy* untuk FTC

Batasan masalah pada Tugas Akhir ini adalah :

- Tidak menangani *preprocessing* data.
- Dataset yang digunakan adalah data yang telah berlabel.
- *Term* diasumsikan terdiri dari sebuah kata bukan phrase.
- Data dokumen berita berbahasa Indonesia yang digunakan oleh penulis adalah dokumen berita *offline*.
- Metode HFTC tidak dibandingkan dengan metode lain, sedangkan FTC dibandingkan dengan salah satu metode klasterisasi konvensional yaitu Kmeans dari tools Weka.

1.3 Tujuan

Tujuan Tugas Akhir ini adalah :

1. Membangun perangkat lunak yang dapat melakukan pengklasteran dokumen berita berbahasa Indonesia dan kelompok hasil pengklasteran tersebut memiliki deskripsi yang mudah dipahami berdasarkan termnya.
2. Melakukan analisis performansi sistem untuk mengetahui kualitas klaster hasil pengklasteran dengan menggunakan F-measure pada HFTC dan *Entropy* pada FTC.

1.4 Metodologi penyelesaian masalah

1. Studi literatur
   Mengumpulkan informasi dan referensi dari buku, majalah, artikel maupun internet yang akan digunakan sebagai teori dasar penyusunan Tugas Akhir yang berkaitan dengan *selection* data, *preprocessing* data dan pengklasteran, serta mengetahui kualitas klaster hasil pengklasteran dengan F-Measure dan *Entropy*.
2. Pengumpulan data
   Mengumpulkan data berupa dokumen teks yang dibutuhkan untuk keperluan proses implementasi dan pengujian metode yang digunakan.
3. Pengembangan perangkat lunak yang meliputi:
   1. Analisa dan Perancangan
      Melakukan analisa penerapan metode yang digunakan dan perancangan akan sistem yang akan diimplementasikan.
   2. Pengkodean
      Mengimplementasikan perancangan menjadi perangkat lunak.
   3. Pengujian

Perangkat lunak diuji dengan parameter *minimum support* untuk selanjutnya dilakukan klasterisasi dan dihitung performansi dari klaster yang dihasilkan.

4. Analisa hasil
   Melakukan analisa terhadap hasil dari klusterisasi dokumen/artikel web berita berbahasa Indonesia dengan menghitung Fmeasure,Entropy, dan waktu yang dibutuhkan untuk proses klasterisasi.
5. Pembuatan laporan
   Pembuatan laporan Tugas Akhir yang mendokumentasikan tahap-tahap kegiatan dan hasil dalam Tugas Akhir ini.

5. Kesimpulan dan Saran

5.1 Kesimpulan

1. Pada HFTC, F-measure nilainya semakin besar dengan *minimum support* yang semakin kecil, karena *frequent term* yang diproses semakin sedikit. Hal ini menunjukkan kualitas klaster yang terbentuk pun semakin bagus.
2. Pengklasteran pada HFTC tidak dapat menggunakan data dalam jumlah besar.
3. Nilai *Entropy* yang dihasilkan pada FTC bervariasi dan tidak memiliki pola pada tiap *minimum support* yang diinputkan.
4. Nilai *Entropy* FTC pada artikel yang berjumlah 45 dan 60 lebih kecil bila dibandingkan dengan nilai *Entropy* rata-rata pada kmeans pada jumlah klaster yang semakin kecil.
5. Hasil pengujian pada FTC menciptakan klaster yang lebih mudah dimengerti deskripsi klasternya dibanding hasil klaster pada Kmeans.
6. Waktu yang dibutuhkan dalam pembentukan klaster akan semakin sedikit seiring dengan makin besanya nilai *minimum support*, karena *term* yang diproses pun semakin sedikit jumlahnya.

5.2 Saran

1. Diperlukan suatu teknik penentuan *minimum support* yang benar-benar tepat untuk mendapatkan *term* yang *frequent*.
2. Diperlukan penentuan minimum support yang tepat untuk menghasilkan klaster yang terbaik.
3. Diperlukan suatu algoritma perbaruan agar dapat menangani data yang lebih besar dengan komputasi waktu yang kecil.
4. Perlu dilakukan perbaikan aplikasi dengan algoritma yang tepat yang dapat melakukan pengambilan data secara *online* dari website.
5. Diperlukan adanya teknik yang dapat menghilangkan overlap pada HFTC sehingga akan lebih efisien dalam penyimpanan data.

Daftar Pustaka

[1] Andreas Hotho, Gerd Stumme, "Conceptual Clustering of Text Clusters". Institute of Applied Informatics and Formal Description Methods AIFB, University of Karlsruhe, German. 2003.
didownload pada tanggal 02 Januari 2008.

[2] Benjamin Chin Ming Fung, Ke Wang, Martin Ester. "Hierarchical Document Clustering".Simon Fraser University, Canada.2003.
didownload pada tanggal 05 Januari 2008.

[3] Bin Tang, Michael Shepherd, Malcolm I. Heywood, Xiao Luo, "Comparing Dimension Reduction Techniques for Document Clustering" Faculty of Computer Science, Dalhousie University, 6050 University Avenue, Halifax, Nova Scotia, Canada.
Didownload pada tanggal 05 Januari 2008.

[4] Christian Borgelt. " Recursion Pruning for the Apriori Algorithm". School of Computer Science, Otto-von-Guericke-University of Magdeburg. German.2004.
Didownload pada tanggal 05 Januari 2008.

[5] Donner, Yoni. 2006. Automated Text Categorization.
Didownload pada tanggal 20 Maret 2008.

[6] Florian Beil, Martin Ester., Xiaowei Xu. "Frequent Term-Based Text Clustering". International knowledge Discovery and Data Mining, KDD'02, Edmonton, Alberta, Canada, 436-442.2002.
Didownload pada tanggal 28 Desember 2007.

[7] Igg Adiwijaya Ph.D., "Text Mining and Knowledge Dsicovery", Kolokium bersama komunitas datamining Indonesia & softcomputing Indonesia.2006.
Didwonload pada tanggal 02 Januari 2008.

[8] J. Han and M. Kamber. *Data Mining : Concepts and Techniques.* Morgan Kaufmann, 2001.

[9] Yudi Wibisono, Masayu Leylia Khodra.:Clustering Berita Berbahasa Indonesia, FPMIPA Universitas Pendidikan Indonesia & Institut Teknologi Bandung.Bandung.2005.
Didownload pada tanggal 10 Januari 2008.

[10] Sebastiani, Fabrizio. 2002. Machine Learning in Automated Text Categorization. ACM Computing Surveys.
Didownload pada tanggal 10 Januari 2008.

[11] http://en.wikipedia.org/wiki/Text_mining
didownload pada tanggal 12 Maret 2008.

[12] http://en.wikipedia.org/wiki/Dimensionality_reduction
didownload pada tanggal 12 Maret 2008.

[13] http://www.indodm.cgi.htm
Tanggal akses: 27 Maret 2008